

Cross-docking benchmark for automated pose and ranking prediction of ligand binding

Shayne D. Wierbowski¹ | Bentley M. Wingert²  | Jim Zheng³ | Carlos J. Camacho²

¹Department of Biology, University of Scranton, Scranton, Pennsylvania

²Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, Pennsylvania

³Department of Computer Science, University of Pittsburgh, Pittsburgh, Pennsylvania

Correspondence

Carlos J. Camacho, 3064 Biomedical Science Tower 3, 3501 Fifth Avenue, Pittsburgh, PA 15260.
Email: ccamacho@pitt.edu

Funding information

National Institutes of Health, Grant/Award Numbers: GM097082, T32EB009403; Department of Defense in partnership with the NSF REU program; National Science Foundation, Grant/Award Number: DBI-1263020

Abstract

Significant efforts have been devoted in the last decade to improving molecular docking techniques to predict both accurate binding poses and ranking affinities. Some shortcomings in the field are the limited number of standard methods for measuring docking success and the availability of widely accepted standard data sets for use as benchmarks in comparing different docking algorithms throughout the field. In order to address these issues, we have created a Cross-Docking Benchmark server. The server is a versatile cross-docking data set containing 4,399 protein-ligand complexes across 95 protein targets intended to serve as benchmark set and gold standard for state-of-the-art pose and ranking prediction in easy, medium, hard, or very hard docking targets. The benchmark along with a customizable cross-docking data set generation tool is available at <http://disco.csb.pitt.edu>. We further demonstrate the potential uses of the server in questions outside of basic benchmarking such as the selection of the ideal docking reference structure.

KEYWORDS

affinity ranking, cross-docking, docking, drug discovery, pose prediction, small molecule, virtual screening

1 | INTRODUCTION

Despite increasing acceptance and utilization of molecular docking toward problems such as de novo drug discovery and lead optimization, significant shortcomings exist in the assessment of docking success, particularly in the lack of unified standards.^{1,2} The lack of a widely accepted standard for calculating docking successes and the limited number of available benchmarking data sets to accurately compare different techniques has led to confusion in the field when it comes to determining success. While resources such as the Community Structure-Activity Resource (CSAR)^{3–5} and Drug Design Data Resource (D3R)^{6–9} have begun to address issues per-

taining to the lack of high-quality training data set, there is still much room for improvement.^{3,4}

When it comes to determining the success of a docking algorithm on a particular data set, several reasonable approaches may be employed. The ability to properly reconstruct a ligand's known binding position may be one of the most highly accepted measures of docking success.^{10,11} Alternatively, the accurate prediction of binding affinity may be selected as the criterion for dock assessment; however, the affinity may only need to be accurate enough to differentiate known binders from known decoys.^{2,10,11} Identification of the original binding receptor when ligands are docked to different proteins may be another measure.

While some data sets exist to fit these means of assessment, there are not yet sufficiently scoped data sets for

Shayne D. Wierbowski and Bentley M. Wingert contributed equally.

each of them. The Database of Useful Decoys-Enhanced (DUD-E) serves as one such standardized data set.² However, DUD-E is specialized to retrospective enrichment of known versus decoy binders and is not suited for correct binding conformation prediction studies. The Astex Non-native set¹² was created to focus on cross-docking. However, it was curated from existing crystal structures on the Protein Data Bank (PDB)¹³ as of 2008 and it is outdated (e.g., very few DUD-E targets). Redocking refers to the docking of a ligand to its *holo* crystal structure. In redocking, a ligand is extracted from a protein-ligand crystal and docked to the same protein receptor. These are significantly easier cases to solve due to the fact that the receptor structure is in its optimal conformation. Cross-docking extracts a ligand from a co-crystal but docks it to another conformation of the same protein rather than the ligand's *holo* structure. Thus, the problem is significantly more challenging and more analogous to the type of *de novo* binding predictions that molecular docking algorithms are intended for.

In situations where cross-docking is properly carried out, two problems emerge. Either individual labs must generate their own cross-docking data sets *ad hoc*¹⁰ or they must rely on the sparse data sets that are available.^{3,4} The generation of *ad hoc* cross-docking data sets proves to be a tedious task and provides no meaningful comparison to studies carried out by others in the field. Reliance on pre-existing data sets falls short because they often do not encompass a wide enough range of proteins or ligand to provide a reliable assessment. Furthermore, these data sets often require a number of processing steps by the user before docking can be carried out.

To this end, we set about to create a new Cross-Docking Benchmark server *that hopefully will be used to generate* a suitable gold-standard to compare different methodologies in a reproducible manner. The benchmark includes a subset of targets from the DUD-E data set and encompasses 4,399 ligand structures for docking. Efforts have been made to make the server immediately ready for docking upon download in order to provide the smoothest docking experience. This curated data set is suitable for use when performing new evaluations of tools and workflows and also provides a quality baseline of docking performance to a rigid receptor for comparison. The automated docking and ranking strategies used to generate these poses have been validated in community-wide prospective evaluations and utilize only publicly available tools^{5,9,14} that have consistently predicted top-of-the-line results for both pose prediction and affinity ranking. We also provide the workflow used to create the benchmark in order to facilitate the creation of cross-docking data sets for any target not included herein. Both the server and the

cross-docking generation tools are freely available at <http://disco.csb.pitt.edu>.

Previous studies have shown that one of the biggest factors in the success of a molecular docking prediction is the choice of the ideal receptor structure to use as the docking reference.^{5, 9, 14} Here, we address this problem using the broad set of targets in the Cross-Docking Benchmark to study how different methods of choosing the reference receptor structure affect overall docking success. Specifically, we addressed previous findings that suggested a partial correlation between the binding pocket volume of a receptor and its effectiveness as a docking reference.⁵ We compare the overall success under the average case, selection based on DUD-E reference, selection based on pocket volume, and picking the best available receptor as identified by our workflow.

2 | MATERIALS AND METHODS

2.1 | Selecting the cross-docking benchmark set of targets

The benchmark was designed around a subset of the targets described in DUD-E.² This set has been broadly used in similar molecular docking problems, it consists on a functionally diverse set of protein targets—including kinases, proteases, signaling receptors, and other enzymes—and thus it is appropriate for our uses. DUD-E provides, as a representative of each target, a single X-ray structure and cocrystallized ligand carefully selected for docking, each of which was downloaded and used to seed our benchmark generation algorithm as described below.

2.2 | Benchmark generation algorithm

Generation of the data set was completed using custom python scripts. The goal was to not only provide a set of PDB structures that could be used in cross-docking experiments, but to process these structures so that they would be docking ready. The final set is available for free download at <http://disco.csb.pitt.edu>. Although it was generated automatically, manual curation of the data was performed to ensure correct processing.

In order to identify the set of structures relevant for each target, the reference structure was used to search for the set of 90% homologous structures using the RSCB PDB's RESTful sequence cluster service.¹³ Each candidate structure was parsed to determine what ligands it contained, and ligand affinity data (i.e., IC50 values used for ranking prediction) for each ligand was obtained where available by consulting The Binding Database,¹⁵ The

PDBBind Database,^{16,17} and Binding MOAD.^{18,19} Those structures which contained at least one ligand were downloaded and separated into distinct protein and ligand(s) files using the PyMol²⁰ API (Schrödinger).

In order to process these structures, and prepare them for docking, it was necessary to align each structure to the reference and to identify the ligand that should be used in docking. Using PyMol, each chain of the candidate structure was successively aligned to the reference protein to identify the chain that both best aligned to the reference and placed a candidate ligand near the reference ligand. If alignment was not possible—either because the root-mean-square deviation (RMSD) of the protein alignment was greater than 4.0 Å or because no candidate ligand was within 4.0 Å of the reference ligand—the structure was removed because these misaligned structures would not provide accurate “known” ligand positions for determination of RMSD of predicted docked poses. Structures in which multiple ligands appeared near the binding pocket—within 5.0 Å of the selected candidate ligand—were also removed because the stable binding in these structures may rely on ligand–ligand interactions that would not be represented in molecular docking predictions between the protein and single ligand. To ensure most efficient docking on the user's end, once the candidate ligand that best fit the reference binding pocket was identified, the candidate protein structure was trimmed to only those chains directly interacting with the ligand—determined as those chains within 10.0 Å of the ligand. From the trimmed protein, the other ligand cofactors and crystal waters within 5.0 Å were saved separately.

The above methodology results in a highly curated set of docking ready protein and ligand structures. The benchmark set is presented as a collection of directories representing each target. Within each target directory, PDB files containing the protein, ligand, other molecules, and water molecules for each structure are retained in a subdirectory. Additionally, four informational logs are presented. The “pdbbs_kept.txt” log contains a list of the PDB id for the final set of structures. The “pdbbs_considered.txt” log contains a list of those PDB IDs which were removed from the set and the reason for their removal. The “lig-map2.txt” log contains a map of PDB ID to ligand ID since for future convenience, the ligand identifiers were all renamed to “LIG.” The “lig_affinity.txt” contains the available ligand binding affinity data available for each ligand.

2.3 | Using benchmark generation to a customizable script

The scripts used to generate our server are provided as a generalized tool allowing users to define their own

docking data sets around specified targets of interest. This custom generation tool is available at <http://disco.csb.pitt.edu/Generate.php>. The generation methodology is as described above, except that candidate structure are selected from user provided seed receptor PDB structure and specified ligand identifier. The user specified PDB structure is downloaded from the RSCB PDB and the specified ligand is extracted. If no specific chains are indicated in the user input, the first instance of the ligand molecule is extracted, and the protein structure is trimmed to only those chains within 10.0 Å. If specific chains are indicated, the ligand and or protein chains are extracted and additional protein chains interacting with the ligand are added as necessary. Results are emailed to the user as soon as they are available.

2.4 | The cross-docking benchmark

In order to establish a fixed cross-docking standard success rate to be used as the benchmark to “beat,” cross-docking predictions were made for all targets using smina¹¹ with default settings. The selection of reference protein structure to cross-dock to was done by selecting the structure that provided the maximal docking efficiency as previously described.^{5,9,14} From smina's cross-docking predictions, the RMSD between the “known” position of the ligand and each of the predicted poses was calculated using Open Babel's root-mean-square method.²¹ Docking success rate was determined as the percentage of predicted dock poses less than 2.0 Å, and statistics is provided for the best Vina scored prediction,²² and for the lowest RMSD pose within the best five ranked predictions. A final benchmark for each target is available for download. Based on these results, each target was classified as an easy, medium, hard, or very hard docking target (threshold at >75%, >50%, >25%, and <25%, respectively).

2.5 | Pocket volume calculation

Pocket volume calculations were carried out using fdpocket.^{23,24}

3 | RESULTS

We present here a Cross-Docking Benchmark intended to serve as a useful tool in molecular docking analyses. The server was created by starting from selected targets in DUD-E,² which was created for testing docking methods. From these 102 DUD-E references, PDB

structures homologous to each target were identified, screened for inclusion of a ligand compound, aligned to the DUD-E reference, and processed to prepare the structures for molecular docking. The benchmark consists of 95 of the DUD-E targets and a total of 4,399 ligands for docking, with an average of 46 ligands available in each target. A breakdown of the targets included is shown in Table S1. The Targets landing page is shown in Figure 1.

The construction of this data set provides a fast and easy way to assess molecular docking algorithms against a fixed standard. Whereas previously these standard docking sets were few and far between or impractically small; Our benchmark provides a simple, preprocessed docking experience. As a gold standard for cross-docking and ranking, a comprehensive docking across every protein structure for each target was performed and the

overall sampling success rate for each was determined. Cross-docking is reported using the optimal receptor structure for each target. The optimal receptor for the pose prediction was the structure that obtained the larger number of ≤ 2.0 Å RMSD to the known position of the ligand. Similarly, the optimal receptor for ranking affinity predictions was the structure that had the best Spearman correlation between Vina score²² function and IC50's for each target. The average RMSD of the best ligand pose for each target was also determined. These benchmark statistics are provided per target on the website. We found that using the aforementioned thresholds, there are 36 Easy, 44 Medium, 10 Hard, and 5 Very Hard targets using Top five pose prediction. When looking at top 1 pose prediction, we found 15 Easy, 52 Medium, 9 Hard, and 19 Very Hard targets. Additionally, for targets with

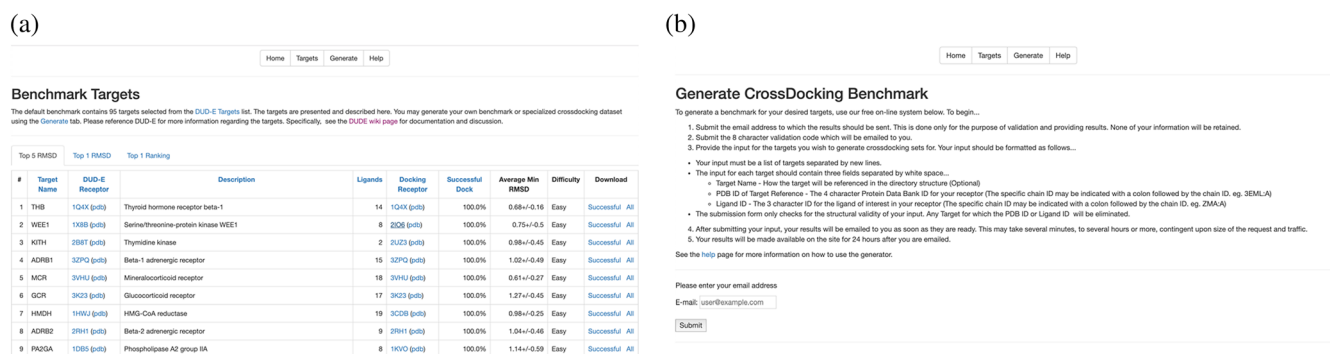


FIGURE 1 Shown are screenshots of the two main functional pages of the Cross-Docking Benchmark webserver. (a) Landing view of the Targets page. Split into results of docking and evaluated by: % successful pose in top 5 poses (top 5 RMSD), % successful pose in top 1 pose (top 1 RMSD), and Spearman correlation of predicted versus experimental affinity values (top 1 ranking). (b) Landing view of Generate page. This is where users can submit a cocrystal structure and have the server return homologous structures with small molecules in the same binding pocket for their own uses. RMSD, root-mean-square deviation

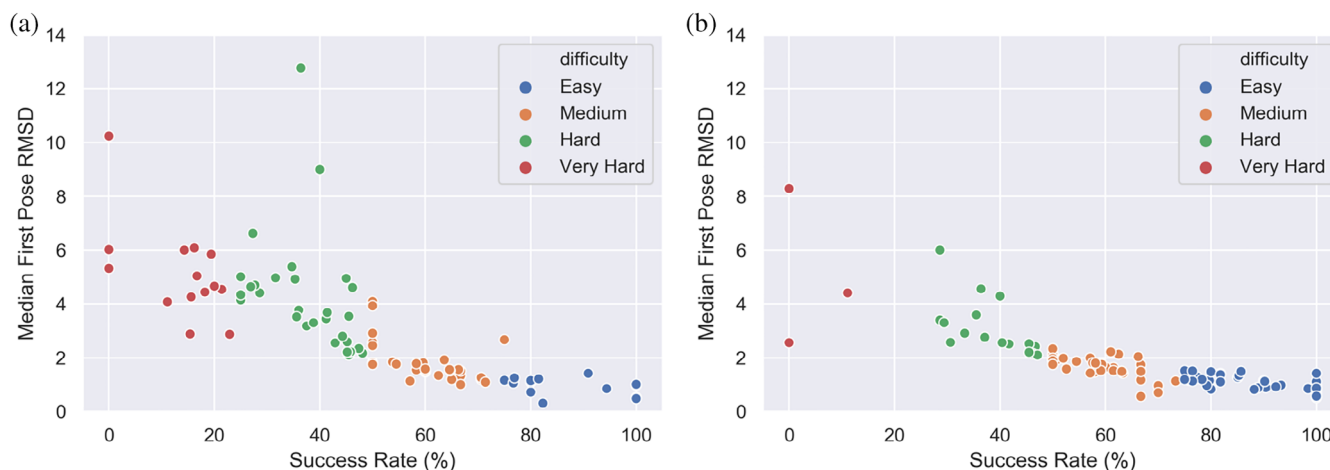


FIGURE 2 Comparison of two measures of docking success; the percentage of compounds with successful pose predictions (RMSD < 2 Å) on the x axis, and the median RMSD in Å of the first pose generated by smina on the y axis. Each dot represents one of the targets in the Benchmark database and is colored by their assigned difficulty label. (a) Top 1 pose. (b) Best of top 5 generated poses. RMSD, root-mean-square deviation

ligands with experimental affinity data, we found a median Spearman correlation of 0.48, which would be at or near the top of past D3R affinity ranking challenges.^{7,8}

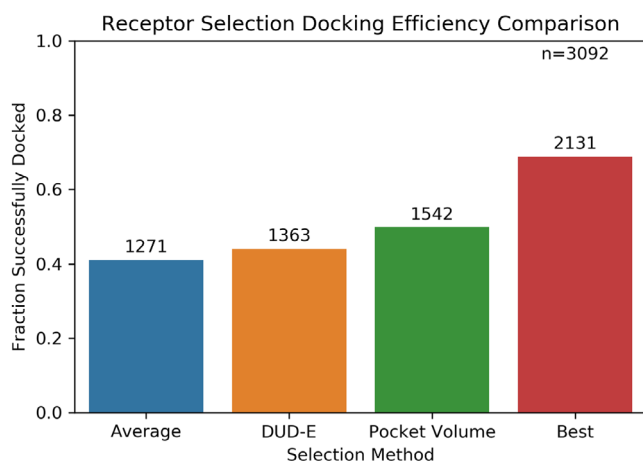


FIGURE 3 Comparing various methods of selecting the receptor structure in a cross-docking prediction. Four methods, random average (41.1%), DUD-E reference (44.1%), largest pocket volume (49.9%), and best (68.9%) were compared for overall docking performance based on sampling of low RMSD poses. Significant improvements were shown when selecting the best receptor compared with the expected average from random selection. These findings demonstrate the importance of proper receptor selection prior to docking. RMSD, root-mean-square deviation

The relationship between a target's median first pose RMSD for all compounds and its amenity to overall success at docking is shown in Figure 2. While both are relevant measures of a target's amenity to docking studies, we believe a target's success rate is the more meaningful (and less prone to outliers) metric. Hence, we use it for our categorization of target docking difficulty.

The docking results indicated that the selection of the correct receptor can have a significant impact on docking performance. Specifically, comparisons were made between overall docking performance under four receptor selection criteria; (a) random selection, (b) using the DUD-E receptor, (c) using the receptor with the greatest binding pocket volume, and (d) using the actual best receptor. Results are presented in Figure 3. The discrepancy between random receptor selection and best receptor selection of nearly 30% highlights the importance of proper receptor selection.

As previously shown,⁵ there seems to be a correlation between pocket volume in a receptor and the ease of docking. The receptor selection based on pocket volume demonstrates nearly a 10% increase in overall docking performance, suggesting that pocket volume could be a reasonable measure for selecting the receptor with better than average performance.

In order to further evaluate the significance of pocket volume in identifying the best receptor structure, pocket volume was plotted against the overall sampling success rate for each receptor. As Figure 4 demonstrates, although

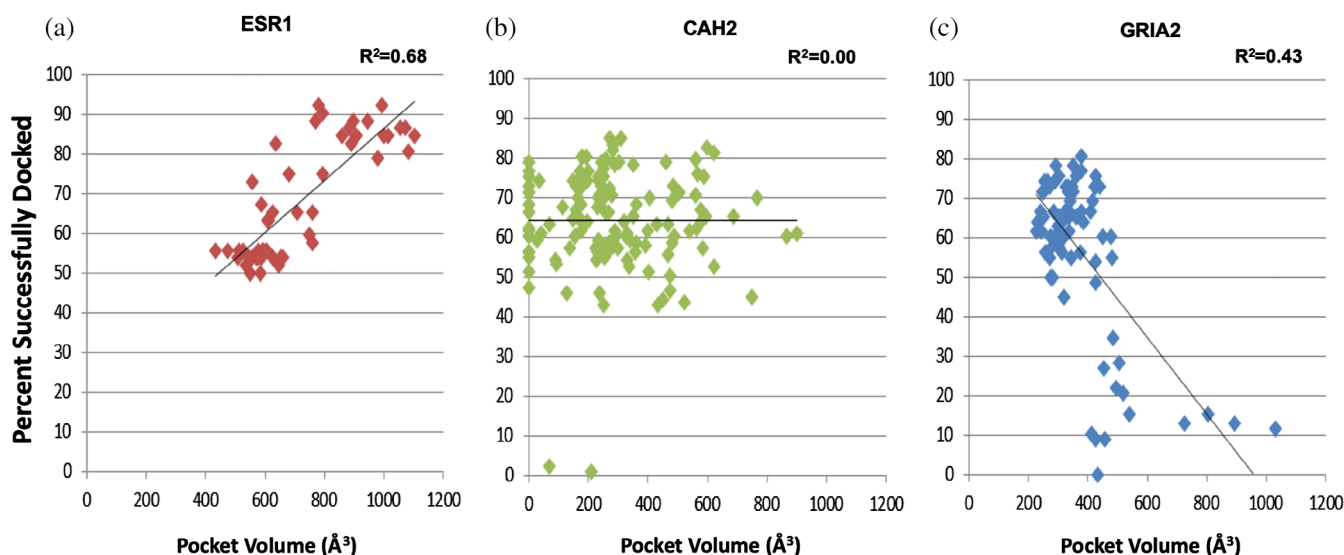


FIGURE 4 Consideration of the correlation between pocket volume of the receptor and its success in cross-docking. Although selection of receptor based on pocket volume produced above average performance, it is clearly not the case that pocket volume serves as a comprehensive indicator of a receptors performance as a docking reference. While targets such as ESR1 (estrogen receptor) (a) have a strong positive correlation, there are also targets such as GRIA2 (glutamate receptor 2) (c) that have a weak to strong negative correlation. Most targets are like CAH2 (carbonic anhydrase II) (b) and exhibit either no or an extremely weak correlation between pocket volume and docking success

the pocket volume can in some instances be a strong indicator of the success of docking, this is far from a rule. Some targets such as ESR1 have strong positive correlation between pocket volume and sampling success ($R^2 = 0.68$), while other targets have weak or even negative correlation. Indeed, the overall correlation between pocket volume and sampling success was found to be weakly negative if anything. However, the fact that some receptors can easily be docked to due to large pocket volume while others cannot suggests that a number of factors play a role in the ease of docking to a receptor and that pocket volume may nonetheless be one of them to consider.

We also investigated if receptors best suited for pose prediction were also well suited for predicting affinity ranking. It would seem intuitive that the two would be related, though in previous blinded docking studies^{7,8,14} this has been observed to not be the case in a small number of targets which in many cases were structurally similar (i.e., kinases). With the larger, more diverse set of targets in the Cross-Docking Benchmark database, we can see that successful pose prediction and successful ligand affinity ranking do not seem to be correlated (Figure 5). This seems to validate our earlier conclusion^{14,25} as well as emphasize the observation that receptor choice is a vitally important part of a successful drug

discovery pipeline and should be undertaken separately for different goals.

4 | DISCUSSION

One of the major requirements in the assessment of various approaches to molecular docking is the existence of a standardized benchmark to compare docking successes. However, multiple criteria for judging the success of molecular docking exist, and there are not yet standard benchmarks representing all of them. For example, DUD-E presents a thorough data set to be used as a benchmark for molecular docking based on enrichment in retrospective recall of known binders compared to known decoys.² While DUD-E focuses on evaluating accurate prediction of binding affinity namely accurate scoring, an equally important component in molecular docking is the ability to accurately generate predicted poses representative of the known binding position namely sampling. The Astex Non-native set,¹² a currently available data set similarly built to focus on cross-docking evaluation, is limited in scale (Table 1). Of note, since ranking is a major challenge in small molecule docking, having an average of three times more structures/ligands per target makes the overall success rates a poor comparison between the two sets.

To this end, we present our Benchmark, as a sizable data set for easy cross-docking testing. The webserver encompasses 95 unique protein targets averaging 46 ligands per target, resulting in 4,399 highly curated structures immediately ready for docking. Furthermore, these structures have each been meaningfully processed in order to facilitate the docking experience and calculation of RMSD between the known and predicted poses. Before this effort, assessment of the success of molecular docking algorithms based on pose recall required either; (a) reliance on significantly limited data sets for cross-docking, (b) reliance on redocking studies, or (c) manual generation of a cross-docking data set specific for in house use. Each of these solutions leaves much to be desired. Contrary to other sets, here we provide the

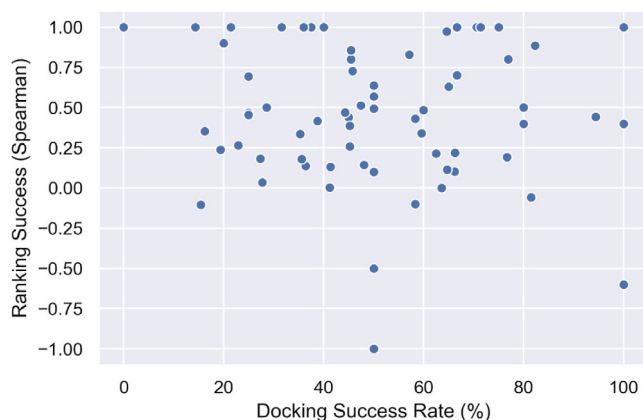


FIGURE 5 Comparison between a target's ability to generate successful pose predictions (x axis) and successfully rank compounds by predicted affinity (y axis)

TABLE 1 Comparison of Cross-Docking Benchmark & existing benchmark set

| Data set | Number of targets | Number of structures | Average # structures per target | Median # of structures per target | Average first pose success rate | Average best pose success rate |
|-------------------------|-------------------|----------------------|---------------------------------|-----------------------------------|---------------------------------|--------------------------------|
| Cross-docking benchmark | 95 | 4,399 | 46 | 19 | 50% | 65%* |
| Astex | 65 | 1,112 | 17 | 6 | 61% | 72% |

Note: High-level comparison between Cross-Docking Benchmark data set and Astex Non-native set. Success rate calculated as described in Verdonk et al.¹² Our Benchmark focuses on top 1 or top 5 subsets of generated poses as opposed to looking at every generated pose, requiring docking programs to surface high-quality poses near the top.

largest set of available targets to date, in a standardized format that is ready for immediate docking with tools used elsewhere in the field. The docking data set was generated by automated methods with a rigid receptor structure and using freely available tools. The methods used have been previously described and shown to be top-of-the-line in community-wide prospective docking challenges.^{5,9,14}

The use of the DUD-E database as the source for targets ensures that a wide range of protein families are included in the data set for cross-docking analysis. And, using our customizable cross-docking data set generation script one can easily generate additional target reference structures. This feature will also allow for quick and easy updating of the Cross-Docking Benchmark data set with new targets of interest to researchers. The server provides a solution to each of the difficulties in molecular docking analyses based on pose recall. The structure of data as a series of targets containing clearly named and separated protein and ligand PDB files facilitates docking. Finally, the docking performance analysis carried out and fully available on the server provides a clear benchmark for small molecule docking on a rigid receptor. Most importantly, the use of this benchmark in future studies will make possible direct comparison of diverse molecular docking approaches.

Our consideration on the selection of an optimal receptor structure for cross-docking demonstrates the need to carefully consider the choice of receptor in instances where multiple receptor structures are available. Although studies have identified this problem in the past,^{5,9,14} we present here one of the first large scale considerations of the overall effect of receptor selection on cross-docking success (Figure 2). We show a nearly 30% improvement in docking results when the best receptor is correctly identified as opposed to when receptors are selected at random (Figure 3).

If receptor selection has such a significant impact on ease of docking, then the next question becomes the proper identification of which features matter in receptor selection. Some findings have suggested that the binding pocket volume may be one important feature.⁵ Our consideration of pocket volume as a selection criterion is consistent with these findings. However, our larger assessment of pocket volume to docking success correlation on a per target basis demonstrates that it is clearly not the best or only factor involved (Figure 3). We feel that this is an important problem, and data sets like ours could provide a doorway to further approaches to identify the meaningful factors at play.


One feature of webserver that makes it an enticing gold-standard benchmark is the high range in difficulty of docking predictions over its targets. Docking success rates vary from 0 to 100% for both top 1 and top

5 measurements, and ranking correlations vary from −1.0 to 1.0 Spearman ρ . This factor should become critical in future assessments of small-molecule docking workflows. While we expect the Cross-Docking Benchmark to serve as a means of comparing docking algorithms, the bulk of data made available by here will also allow for analysis of the successes and shortcomings of individual algorithms on particular types of structures. Moreover, the existence of an automated gold-standard data set will aid development of automated docking methods which are able to improve upon current best predictions, especially of the difficult “Hard” and “Very Hard” targets.

ACKNOWLEDGMENTS

The work was funded by U.S. National Institutes of Health grant numbers GM097082 and T32EB009403. The TECBio REU @ Pitt provided the research opportunity that made this research possible. Partial funding support was provided by the National Science Foundation under Grant DBI-1263020 and by the Department of Defense in partnership with the NSF REU program. We thank all members of the Camacho lab, particularly Dr. Carlos Camacho, Matthew Baumgartner and Dr. David Koes.

ORCID

Bentley M. Wingert  <https://orcid.org/0000-0003-1000-8498>

REFERENCES

1. Jain AN, Nicholls A. Recommendations for evaluation of computational methods. *J Comput Aided Mol Des.* 2008;22: 133–139.
2. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Database of Useful Decoys, Enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J Med Chem.* 2012;55: 6582–6594.
3. Dunbar JB, Smith RD, Yang CY, et al. CSAR benchmark exercise of 2010: Selection of the protein–ligand complexes. *J Chem Inf Model.* 2011;51:2036–2046.
4. Dunbar JB, Smith RD, Damm-Ganamet KL, et al. CSAR data set release 2012: Ligands, affinities, complexes, and docking decoys. *J Chem Inf Model.* 2013;53:1842–1852.
5. Baumgartner MP, Camacho CJ. Choosing the optimal rigid receptor for docking and scoring in the CSAR 2013/2014 experiment. *J Chem Inf Model.* 2016;56:1004–1012.
6. Gathiaka S, Liu S, Chiu M, et al. D3R grand challenge 2015: Evaluation of protein–ligand pose and affinity predictions. *J Comput Aided Mol Des.* 2016;30:651–668.
7. Gaieb Z, Liu S, Gathiaka S, et al. D3R Grand Challenge 2: Blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J Comput Aided Mol Des.* 2018;32:1–20.
8. Gaieb Z, Parks CD, Chiu M, et al. D3R Grand Challenge 3: Blind prediction of protein–ligand poses and affinity rankings. *J Comput Aided Mol Des.* 2019;33:1–18.

9. Ye Z, Baumgartner MP, Wingert BM, Camacho CJ. Optimal strategies for virtual screening of induced-fit and flexible target in the 2015 D3R Grand Challenge. *J Comput Aided Mol Des*. 2016;30:695–706.
10. Kim R, Skolnick J. Assessment of programs for ligand binding affinity prediction. *J Comput Chem*. 2008;29:1316–1331.
11. Koes DR, Baumgartner MP, Camacho CJ. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J Chem Inf Model*. 2013;53:1893–1904.
12. Verdonk ML, Mortenson PN, Hall RJ, Hartshorn MJ, Murray CW. Protein–ligand docking against non-native protein conformers. *J Chem Inf Model*. 2008;48:2214–2225.
13. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res*. 2000;28:235–242.
14. Wingert BM, Oerlemans R, Camacho CJ. Optimal affinity ranking for automated virtual screening validated in prospective D3R grand challenges. *J Comput Aided Mol Des*. 2018;32:287–297.
15. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: A web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res*. 2007;35:D198–D201.
16. Wang R, Fang X, Lu Y, Wang S. The PDBbind database: Collection of binding affinities for protein–ligand complexes with known three-dimensional structures. *J Med Chem*. 2004;47:2977–2980.
17. Wang R, Fang X, Lu Y, Yang CY, Wang S. The PDBbind database: Methodologies and updates. *J Med Chem*. 2005;48:4111–4119.
18. Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA. Binding MOAD (Mother of All Databases). *Proteins*. 2000;60:333–340.
19. Ahmed A, Smith RD, Clark JJ, Dunbar JB Jr, Carlson HA. Recent improvements to Binding MOAD: A resource for protein–ligand binding affinities and structures. *Nucleic Acids Res*. 2015;43:D465–D469.
20. The PyMOL Molecular Graphics System, Version 1.7.4.4. Schrödinger, LLC
21. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchinson GR. Open Babel: An open chemical toolbox. *J Cheminform*. 2011;3:33.
22. Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J Comput Chem*. 2010;31:455–461.
23. Guilloux VL, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*. 2009;10:168.
24. Schmidtke P, Bidon-Chanal A, Luque J, Barril X. MDpocket: Open source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics*. 2011;27:3276–3285.
25. Wingert BM, Camacho CJ. Improving small molecule virtual screening strategies for the next generation of therapeutics. *Curr Opin Chem Biol*. 2018;44:87–92.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Wierbowski SD, Wingert BM, Zheng J, Camacho CJ. Cross-docking benchmark for automated pose and ranking prediction of ligand binding. *Protein Science*. 2020; 29:298–305. <https://doi.org/10.1002/pro.3784>